DOCUMENT RESUME

ED 293 851 TM 011 396

AUTHOR Haladyna, Thomas M.; Downing, Steven M.
TITLE Functional Distractors: Implications for Test-Item Writing and Test Design.
PUB DATE Apr 88
NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 5-9, 1988).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; Item Analysis; *Multiple Choice Tests; Physicians; Standardized Tests; *Test Construction; Test Format; Test Items; Test Reliability
IDENTIFIERS *Distractors (Testing)

ABSTRACT

The proposition that the optimal number of options in a multiple choice test item is three was examined. The concept of functional distractor, a plausible wrong answer that is negatively discriminating when total test performance is the criterion, is discussed. Three distinct groups of achievers (high, middle, and low) on a national standardized achievement test for physicians were identified. The number of functional distractors was identified for each sample condition and each item. Results suggest, as have theoretical analyses, that more functional distractors are desirable. However, with higher achieving students the number of functional distractors, and consequently the number of options, had no effect. The time and effort devoted to additional option item development is probably not worth the gains in item discrimination and reliability for high and middle achieving examinees. The three-option format is efficient to construct, and it results in better domain-referenced measures of achievement. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED293851

Functional Distractors:

Implications for Test-Item Writing and Test Design [1]

Thomas M. Haladyna
Arizona State University
West Campus

and

Steven M. Downing
National Board of Medical Examiners

U S DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it

☐ Minor changes have been made to improve reproduction quality

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

THOMAS M. HALADYNA

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Paper presented at the annual meeting of the American Educational Research Association, New Orleans, Louisiana

[1]Much of this research was carried out while the second author was at the American College Testing Program (ACT), Iowa City, Iowa.

TM 011 396

ABSTRACT

Functional Distractors:

Implications for Test-Item Writing and Test Design

A recent review of research on the desirable number of options for a multiple-choice test item reveals that two or three options may be suitable for most examinees for an achievement test. Most textbooks recommend four- or five- option items. The number of options in an item should be based upon the functionality of each option. In this study, a functional distractor is defined, research on the optimal number of options is reviewed, and a study is reported on the number of functional distractors in a high quality achievement testing program. This research examines the proposition that the number of functional distractors per item is optimally around two, and that the three-option format is not only more efficient to construct but also leads to better domain-referenced measures of achievement.

Functional Distractors:

Implications for Test-Item Writing and Test Design

The design of any multiple-choice test item is often guided by one's experiences, common sense, and item-writing lore, passed on from mentors or textbooks. In a review of 46 textbooks treating the topic of writing multiple-choice test items, Haladyna and Downing (in press, a) reported that authors often disagreed in their advice on the ideal number of options. Some recommended four or five, while others recommended producing as many as were feasible or plausible.

The present study tested the proposition that the optimal number of options is three. In most well designed achievement tests, more than two functional distractors are rare and do not necessarily contribute to more effective measurement of achievement. Given that this proposition holds, test developers at all levels and in all areas would be better served to design test items which contain fewer, but more functional, distractors.

The benefits of using fewer distractors is to reduce item development time, reduce the length of tests, reduce reading and administration time, and still retain the measurement properties desired. Additionally, the use of functional distractors enable the productive use of promising new technologies, such as polychotomous scoring models (Bock, 1972; Thissen, 1975; Sympson, 1985) to provide better estimates of achievement or ability.

The proposition that three options is optimal for most testing purposes derives from an analysis of past and current research on this topic from both theoretical and empirical perspectives.

Research on the Optimal Number of Options

Theoretical Perspectives. Lord (1944) conducted one of the earliest studies; he developed a formula for predicting changes in reliability as a function of the number of options added to any multiple-choice item. Lord's study suggested that a three-option item is optimal for most examinees. Tversky (1964) reached the same conclusion, based on an analysis of three criteria (discriminability, power, and information of a test). Studies by Ebel (1969), Grier (1975; 1976), and again by Lord (1977) support these findings. Lord's more recent study is most informative about where on the achievement scale the three-option item works best. Using item information curves, Lord (1977) shows that the three-option item provides the most information at the midrange of the score scale, while the two-option item works best at the upper range of the score scale. Four- or five-option items work best at the lower range, where guessing is more frequent and the plausibility of wrong answers is more likely to prove effective. Levine and Drasgow (1983) generally confirm Lord's analysis.

Budescu and Nevo (1985) take issue with the law of proportionality which is often used in these theoretical studies. This law states that the total testing time is proportional to the number of items and options on the test. Their empirical study provides data to refute this law, thus challenging the validity of the conclusion about the optimality of three options. However, their data show that administration time is greater with the use of more options. No one can dispute that writing more distractors takes more time.

Empirical Perspectives. Haladyna and Downing (in press, b) presented a synthesis of 23 studies on the number of options. Unfortunately, many researchers have considered mainly item difficulty and not treated the more important item discrimination, test score reliability, and validity issues related to the choice of the number of options. These studies, limited as they are, show that slight gains in item discrimination are achieved through the use of more options, but these gains are not meaningful. None of these studies tested Lord's analysis that there is differential value for the number of options as a function of the achievement level of the examinee, and none of these studies examined the functionality of distractors as they contribute to measurement or the improvement of items.

Functional Distractors. While this theoretical and empirical research suggests using only three options for a multiple-choice test items, the functionality of distractors has not always been considered in these prior studies. For instance, infrequently selected distractors should be eliminated because only random guessers choose these options. Such non functional distractors can be eliminated on logical or empirical grounds.

A multiple-choice test item is usually designed to satisfy content specifications which are operationalized through a set of objectives, a test blueprint, or a description of competence in a profession. The assembly of items into an achievement test therefore satisfies the bases for content-valid test score interpretations.

The multiple-choice item is traditionally evaluated based on difficulty and discrimination besides content considerations. Classical test theory dictates that the optimal item has moderate difficulty, with item discrimination being as high

as possible. This condition tends to maximize both the variance of scores in the distribution and test score reliability.

Functional distractors are negatively discriminating when total test performance is the criterion, if each distractor is to work as intended. Discrimination is assessed with the point-biserial correlation between performance on each distractor and total test score performance; negative distractors should exhibit negative correlations carefully evaluated with respect to their continued use in the test item. Since distractors are intended to be plausible wrong answers, it seems both illogical and undesirable for such options to have positive relationships to the total test score.

From item response theory perspective, items are evaluated on the basis of unidimensionality or fit to the dimension under consideration. In item response theory, items are also evaluated in terms of an information function, which defines the ability of the item to measure at different levels of achievement (Lord, 1980). Distractors will often display a negative item characteristic curve (Sympson, 1985; Thissen, 1975) which are useful in polychotomous scoring models.

Since the objective of measurement is to reduce error to allow valid interpretations from test data, the study and control of distractors in the framework of either classical or item response theory is justified. Whether scoring is dichotomous or polychotomous, the careful construction and thoughtful evaluation of distractors appears justified.

If the item characteristic curve for the keyed answer is positive, it can be shown that the collective item characteristic curves for wrong options must be negative. This property can be illustrated in the following item:

	OPTIONS					
	A	B	C*	D	E	A+B+D+E
Upper Third	4	0	68	11	17	32
Middle Third	10	2	62	11	15	38
Lower Third	28	3	48	13	8	52

The three performance levels for the correct response, C, mark the approximate item characteristic curve. Figure 1-a illustrates a traditional item characteristic curve. The sum of responses of the distractors (A, B, D, and E) marks an approximate negative item characteristic curve. Distractor A is functional because it displays a negatively sloping item characteristic curve and a negative point-biserial relationship. Figure 1-b illustrates a desirable negatively sloping item characteristic curve.

If an option is not often selected by examinees, one should question that option's usefulness as a distractor. Thus, simple frequency of selection of an option can stand as one criterion to evaluate functionality. Those seldom selected options should be removed, since the length of the item, administration time, and amount of extraneous reading are negative aspects of having useless options in an item. Option B has this characteristic in the above example. Figure 1-c illustrates this condition.

Figure 1-d illustrates another instance a non functional distractor, one where there are sufficient responses to the distractor but no intelligible pattern as a function of student achievement. The point-biserial relationship between performance on this distractor and total test performance is close to zero. Consequently, this distractor adds no information in polychotomous scoring. Option D represents this characteristic in the example above example.

****

Insert Figure 1 here

****

The last example, option E, has a positive item characteristic curve and positive point-biserial correlation with the total score. Given that distractors should have a negative sloping item characteristic curve, this option interferes with the measurement objective by supplying a source of error.

To summarize, a functional distractor has the following characteristics:

1. A significant negative point-biserial relationship to total test score.
2. A negatively sloping item characteristic curve; and
3. A frequency of response greater than 5% for the total group of examinees.

Dysfunctional distractors include all other conditions, and are illustrated in the example. These include options which are (1) infrequently selected, (2) have no statistically significant relationship to the criterion, or (3) have a positive, significant relationship to the criterion.

The Present Study

If the optimal number of options is three, then we would expect several conditions to be present in standardized achievement tests.

1. The most frequently observed number of functional distractors for a representative sample of items should be two. For high achievers this number should be one, and for low achievers this number should be three or four, because low achievers are more likely to guess and thus display a greater tendency to be drawn to plausible distractors.

2. A negative relationship between the number of functional distractors and item difficulty for a heterogeneous sample is predicted, because low

achievers tend to use more options than high achievers. Also there is a natural limit for functional distractors, easy items have distractors which are seldom used, so one would expect such items to exhibit few functional distractors. With the lower achieving students, we would expect a negative or no relationship between the number of functional distractors and difficulty. Faulty items would be likely to have few or no functional distractors.

3. The relationship between the number of functional distractors and item discrimination should be positive with a heterogeneous sample. The more distractors, the more likely the item will discriminate. This premise generally supports the use of more options. However, when examining the usefulness of functional distractors with high achievers, this relationship should not be present. With low achievers, this relationship should be more pronounced, because plausible, functional distractors are more likely to be useful.

4. The mean difficulty and mean discrimination is a function of the number of functional distractors for each item. With difficulty, items with no functional distractors are likely to be easy. Items with one or two functional distractors should be moderate in difficulty, while items with three or four functional distractors should be difficult.

For discrimination, items with no functional distractors are either flawed or very easy and discrimination should be low. Items with two functional distractors should have moderate to high discrimination, while items with three or four functional distractors should have higher discrimination.

However, the analysis based on the entire sample may be misleading since previously reviewed theoretical research predicts that items with two functional distractors are best for high achievers and that three and four functional distractors serve best for middle and low achievers. This study separated three

distinct groups of achievers (high, middle, and low), and an analysis determined in which groups did two, three, or four functional distractors work best.

METHOD

Data Base

The data for this study came from a national standardized achievement test for physicians. This test is part of a continuing education program and is given annually to over 1,000 specialists in a surgical specialty of medicine. The standards for item and test development are quite high, and include such activities as training of item writers, use of detailed content specifications, item review by colleagues, editorial and psychometric reviews, pretesting, and selection of items by a panel of national experts.

Analysis of Data

The sample of 1,111 examinees was divided into three approximately equal groups for the purposes of this study. The number of functional distractors was identified for each sample condition and each item. The criteria for determining distractor functionality were, as stated before, (1) lack of a negatively sloping item characteristic curve, (2) lack of a negative correlation between distractor and test performance, and (3) selection of a response by less than 5% of the examinees. Tabulations were made of the number of items containing none, one, two, three and four functional distractors for each of the four sample conditions. Analyses of variance were done for item difficulty and discrimination based on the number of functional distractors. These were done for each sample condition.

RESULTS AND DISCUSSION

The top of Table 1 provides descriptive statistics from the data base for the total sample and each subsample. The test was moderately difficult, scores

ranged widely, and the overall KR-20 estimate of reliability was .91. The variability of scores for the upper third and middle thirds was considerably more restricted than the lower third for the three sample conditions.

****

Insert Table 1 here

****

The first hypothesis dealt with the frequency of functional distractors in a well developed achievement test. Table 1 provides information about the extent to which these items had functional distractors. For the total sample, there were only 11 items which had four functional distractors, while 73 items had two functional distractors. Based on the criteria for functionality and on the results with this specific test, two functional distractors per item seems to be typical.

When the number of functional distractors is determined from item analyses based on each sample condition, another pattern of results exists For the upper third, one functional distractor per item was most often noted; no items had four functional distractors. For the middle 1/3, the pattern was similar to that of the upper third. For the lower third, one and two functional distractors appears most often, while only one item had four functional distractors.

A restriction in the range of scores will attenuate discrimination, thereby making the detection of functional distractors less likely.
However, if a distractor is to discriminate between high and low achieving examinees at any level, it must be functional. In this study, one or two functional distractors was typical for the lower achieving sample.

Item difficulty. For the total sample, there is a systematic negative relationship between item difficulty and number of functional distractors. As expected, items with no functional distractors are easy, while those with three or four functional distractors are the most difficult. If we use representativeness as the criterion for evaluating items, the items with the mean difficulty closest to overall test score difficulty of .688 were those having two functional distractors. Fewer functional distractors resulted in easier items, more functional distractors resulted in harder items.

Looking at these results within each sample condition, with the upper third group, the most representative set of items had one functional distractor, again supporting Lord's contention that two options were optimal for high achievers (Lord, 1977). Also, with the middle group, one option produced items closest to the average difficulty. With the lower group, it seemed immaterial how many functional distractors there were as each category yielded item difficulties at about the same level as the rest ($F=0.48$, $p=.751$).

Item discrimination. As noted in the lower portion of Table 1, the greater number of functional distractors in the item, the higher the discrimination. The trend is in a positive direction, and the correlation between item discrimination and the number of functional distractors for this sample is .333. This finding supports the proposition that three or four functional distractors are desirable.

However, item discrimination and functionality for the upper 1/3 of the sample, discrimination is nearly uniform. That is, there is no relationship between the number of functional distractors and item discrimination in this upper level sample ($r= -.001$). With the middle 1/3, there is a tendency to favor more distractors, which is sharply increased with the lower 1/3 sample.

Overall, these results show, as others have suggested in theoretical analyses, that more functional distractors are desirable, yet with higher achieving students, the number of functional distractors and hence the number of options has no effect.

CONCLUSIONS

Research on the number of options over the past 60 years has suggested using fewer options, with three being desirable for most measurement purposes. The strongest rationale for this recommendation is higher efficiency, obtained by preparing fewer distractors and administering items which take up less space and require less reading, thus reducing administration time. This study has supported the proposition that fewer options are desirable for some circumstances. More important, while it may be argued that the use of more functional distractors leads to items which are more discriminating, this research has also shown that items with four functional distractors are rare, at least in this sample. Is the time and effort devoted to five-option item development and testing worth the gains in item discrimination and reliability? Probably not with high or middle achieving examinees.

As testing moves toward more adaptive procedures, polychotomous scoring, and wider implementation of item respon theories, the idea of functional distractors should lead test developers to a conclusion that fewer options of higher quality will produce better test scores than five-option items containing, on the average, only two functional distractors. Item writers may wish to develop more options, but the continued use of so many non functional distractors provides no positive advantage over the three-option format with two functional distractors.

Also, there are emerging theories of item development which focus on the functionality of distractors from a logical, judgmental, and theoretical bases (Tatsuoka, 1983; Tatsuoka and Tatsuoka, 1982, 1983; Webb, Herman, Cabello, 1986). Statistical and cognitive learning theories propose integrating teaching and testing to help test designers to build better items, and hence better tests (Roid and Haladyna, 1982). The consideration of the functionality in distractors and application to item design and analysis should contribute to this growing technology for achievement testing.

REFERENCES

1. Bock, k. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 37, 29-51.

2. Budescu, D. V., & Nevo, B. (1985). Optimal number of options: An investigation of the assumption of proportionality. Journal of Educational Measurement, 22, 183-196.

3. Ebel, R. L. (1969). Expected reliability as a function of choices per item. Educational and Psychological Measurement, 29, 565-570.

4. Grier, J. B. (1975). The number of alternatives for optimum test reliability. Journal of Educational Measurement, 12, 109-113.

5. Grier, J. B. (1976). The optimal number of alternatives at a choice point with travel time considered. Journal of Mathematical Psychology, 14, 91-97.

6. Haladyna, T. M. & Downing, S. M. (in press, a) A taxonomy of multiple-choice item-writing rules. Applied Measurement in Education.

7. Haladyna, T. M. & Downing, S. M. (in press, b) The validity of a taxonomy of multiple-choice item-writing rules. Applied Measurement in Education.

8. Levine, M. V. & Drasgow, F. (1983). The relation between incorrect option choice and estimated ability. Educational and Psychological Measurement, 43, 675-685.

9. Lord, F. M. (1944). Reliability of multiple-choice tests as a function of number of choices per item. Journal of Educational Psychology, 35, 175-180.

10. Lord, F. M. (1977). Optimal number of choices per item-A comparison of four approaches. Journal of Educational Measurement, 14, 33-38.

11. Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

12. Roid, G. H. & Haladyna, T. M. (1982). A technology for test-item writing. New York, NY: Academic Press.

13. Sympson, J. B. (1986). Extracting information from wrong answers in computerized adaptive testing. Paper presented in B. F. Green (Chair), New Developments in Computerized Adaptive Testing. Symposium conducted at the annual meeting of the American Psychological Association, Washington, DC.

14. Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. Journal of Educational Measurement, 20, 345-354.

15. Tatsuoka, K. K. & Tatsuoka, M. M. (1982). Detection of aberrant response patterns. Journal of Educational Statistics, 7, 215-231.

16. Tatsuoka, K. K. & Tatsuoka, M. M. (1983). Spotting erroneous rules of operation by the individual consistency index. Journal of Educational Measurement, 20, 221-230.

17. Thissen, D. M. (1976). Information in wrong responses to the Raven Progressive Matrices. Journal of Educational Measurement, 14, 201-214.

18. Tversky, A. (1964). On the optimal number of alternatives at a choice point. Journal of Mathematical Psychology, 1, 386-391.

19. Webb, N. M., Herman, J. L., & Cabello, B. Diagnosing students' errors from their response selections in language arts. Journal of Educational Measurement, 23, 163-170.

Table 1

Descriptive Statistics from the Data Base

Sample Conditions	Sample Size	Mean	Stan. Dev.	Item Diff.	Item Disc.	KR-20
Total	1111	137.6	20.6	.688	.236	.91
Upper 1/3	371	159.0	7.5	.795	.101	.49
Middle 1/3	370	139.6	5.0	.698	.059	--
Lower 1/3	370	114.3	13.8	.571	.147	.77

Distribution of Number of Functional Distractors For Varying Samples

Sample Conditions	Number of Functional Distractors				
	Zero	One	Two	Three	Four
Total	13	49	73	54	11
Upper 1/3	67	90	37	6	0
Middle 1/3	62	85	44	9	0
Lower 1/3	22	70	78	29	1

Analyses of Variance on Item Difficulty and Item Discrimination Using Number of Functional Distractors as the Independent Variable

Sample Conditions	Item Difficulty Number of Functional Distractors					F	p	Effect Size
	Zero	One	Two	Three	Four			
Total	.819	.752	.675	.626	.637	11.24	.001	.187
Upper 1/3	.758	.671	.627	.538	----	13.80	.001	.174
Middle 1/3	.756	.684	.622	.573	----	13.15	.001	.167
Lower 1/3	.704	.700	.677	.676	.760	0.48	.751	----

Sample Conditions	Item Discrimination Number of Functional Distractors					F	p	Effect Size
	Zero	One	Two	Three	Four			
Total	.162	.208	.238	.268	.311	6.20	.001	.113
Upper 1/3	.232	.241	.233	.222	----	0.16	.926	----
Middle 1/3	.200	.240	.264	.307	----	5.76	.001	.081
Lower 1/3	.125	.205	.258	.337	----	23.93	.001	.329

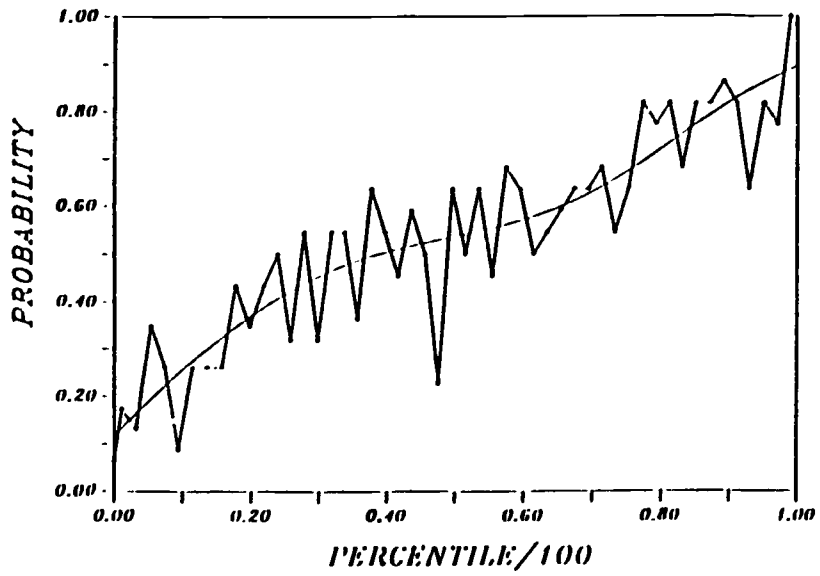

Figure 1-a: Typical item characteristic curve

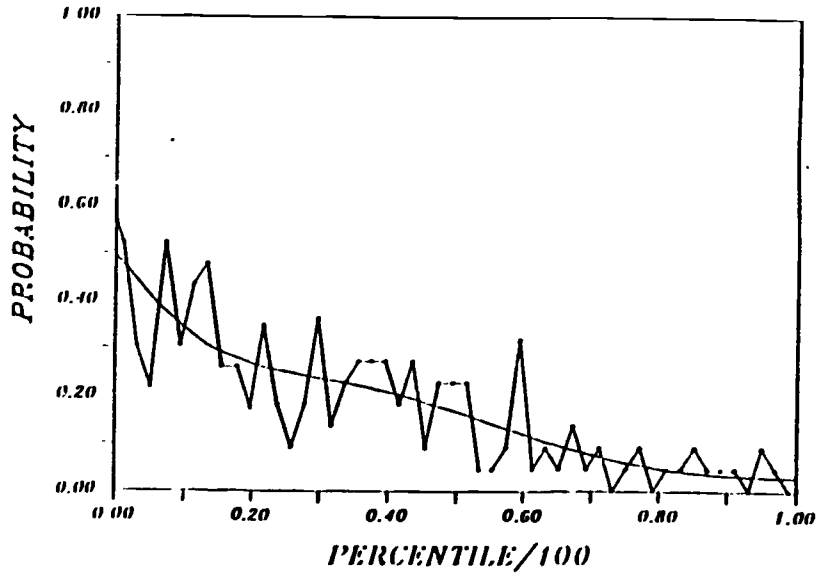

Figure 1-b: Negatively sloping item characteristic curve

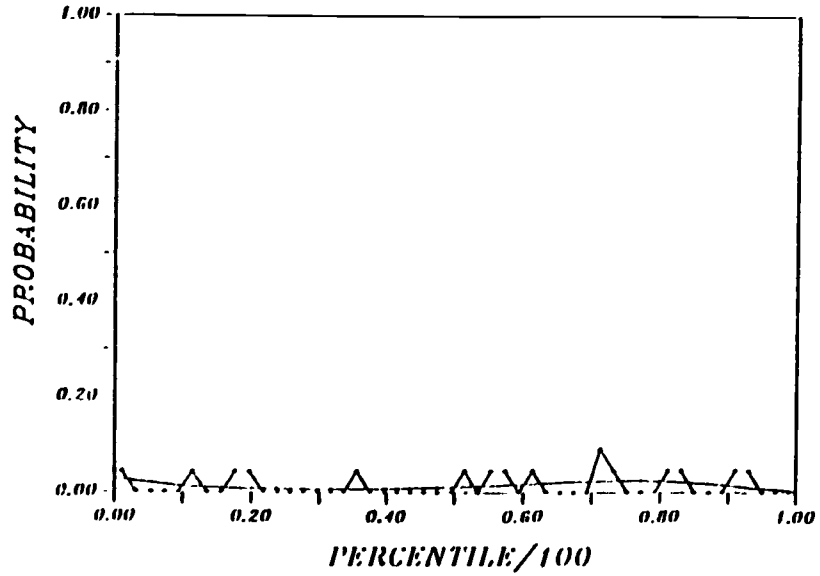

Figure 1-c: Low response item characteristic curve

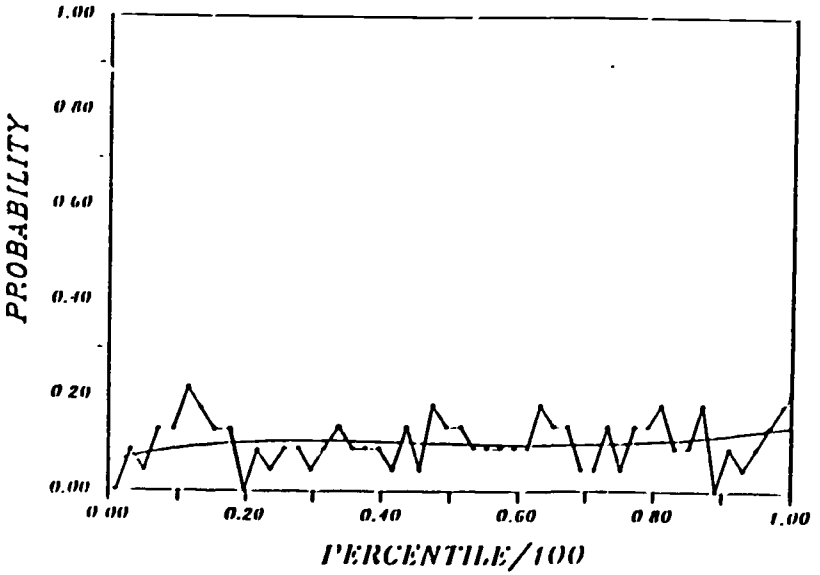

Figure 1-d: Flat item characteristic curve